BETTER MANAGEMENT™

Technical Requirements

- **Access Webcast using Internet Explorer (please disable pop-up blocker)**
- **Program audio available through your computer OR**
- **To listen to audio via your phone:**

  Step 1: Dial the conference access number: 1-866-551-3680 or 1-212-401-6760

  Step 2: Enter PIN code: 5508890#

  Step 3: You will be placed on hold until the event begins

Top 10 Data Mining Mistakes

John F. Elder IV, PhD
Founder and CEO, Elder Research, Inc.

Visionary perspectives for management insights

BETTER MANAGEMENT™

Top 10 Data Mining Mistakes

John F. Elder IV, PhD
Founder and CEO, Elder Research, Inc.

Visionary perspectives for management insights

John F. Elder IV, PhD
Founder and CEO, Elder Research, Inc.

www.datamininglab.com

Top 10 Data Mining Mistakes
-- and how to avoid them

BetterManagement Audio Seminar
August 27, 2009

John F. Elder IV, Ph.D.
elder@datamininglab.com

Elder Research, Inc.
300 West Main Street, Suite 301
Charlottesville, Virginia 22903
434-973-7673
www.datamininglab.com

Elder Research, Inc.
Data Mining & Pattern Discovery

You've made a mistake if you…

1. Lack Data
2. Focus on Training
3. Rely on One Technique
4. Ask the Wrong Question
5. Listen (only) to the Data
6. Accept Leaks from the Future
7. Discount Pesky Cases
8. Extrapolate
9. Answer Every Inquiry
10. Believe the Best Model

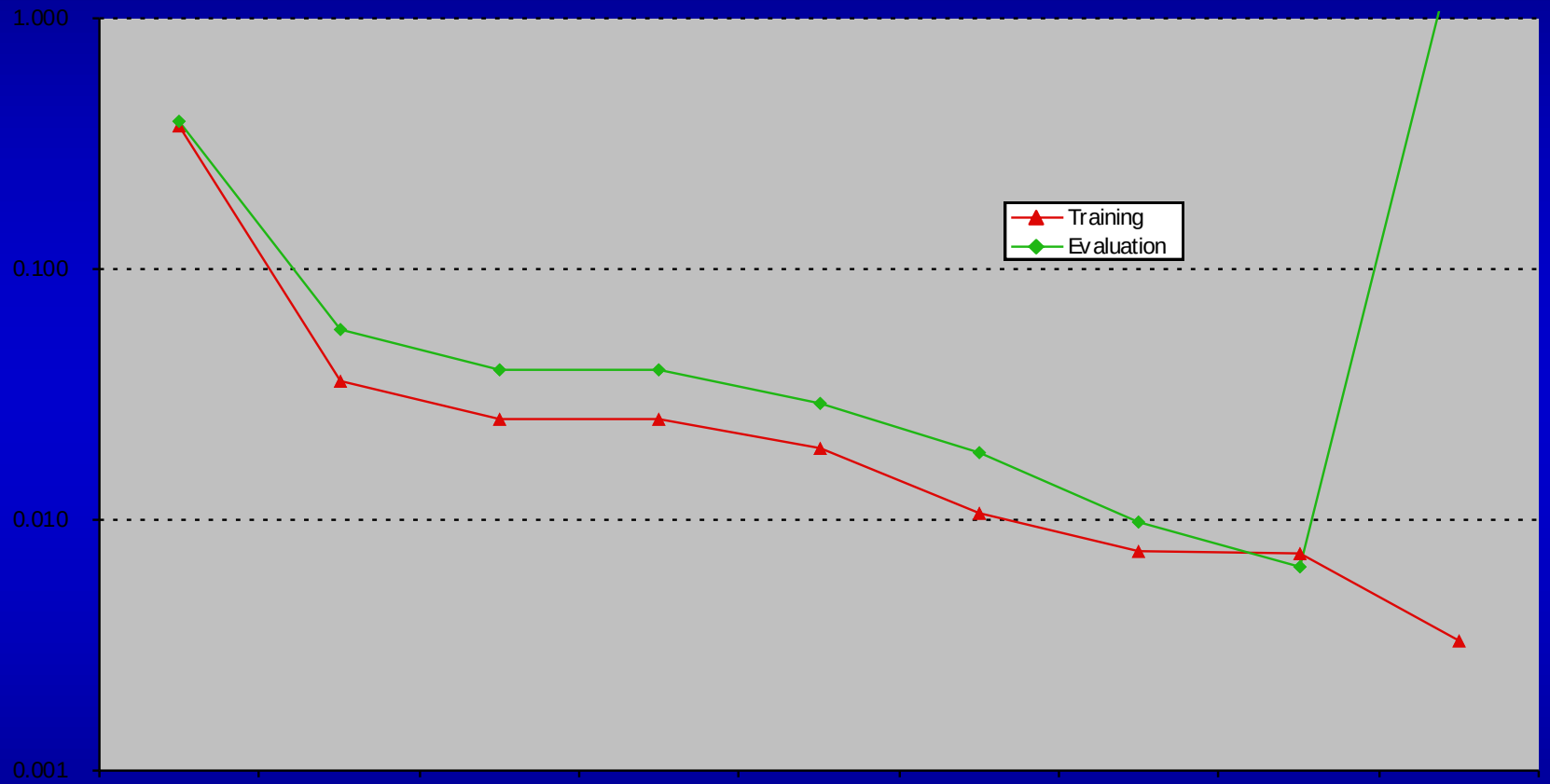
1. Lack Data

- Need labeled cases for best gains to classify or estimate; *clustering* is much less effective. Interesting known cases may be exceedingly rare.
- Ex: Fraud Detection (Government contracting): Millions of transactions, a handful of known fraud cases. Many fraud cases likely mislabeled clean. Only modest results initially.
- Ex: Fraud Detection (Taxes; collusion): Surprisingly many known cases -> stronger, immediate results.
- Ex: Credit Scoring: Capital One (randomly) gave credit to thousands of applicants who were risky by the conventional scoring method, and monitored them for two years. Then, estimated risk using what was known at start. This investment in *creating* the right kind of data paid off.

2. Focus on Training

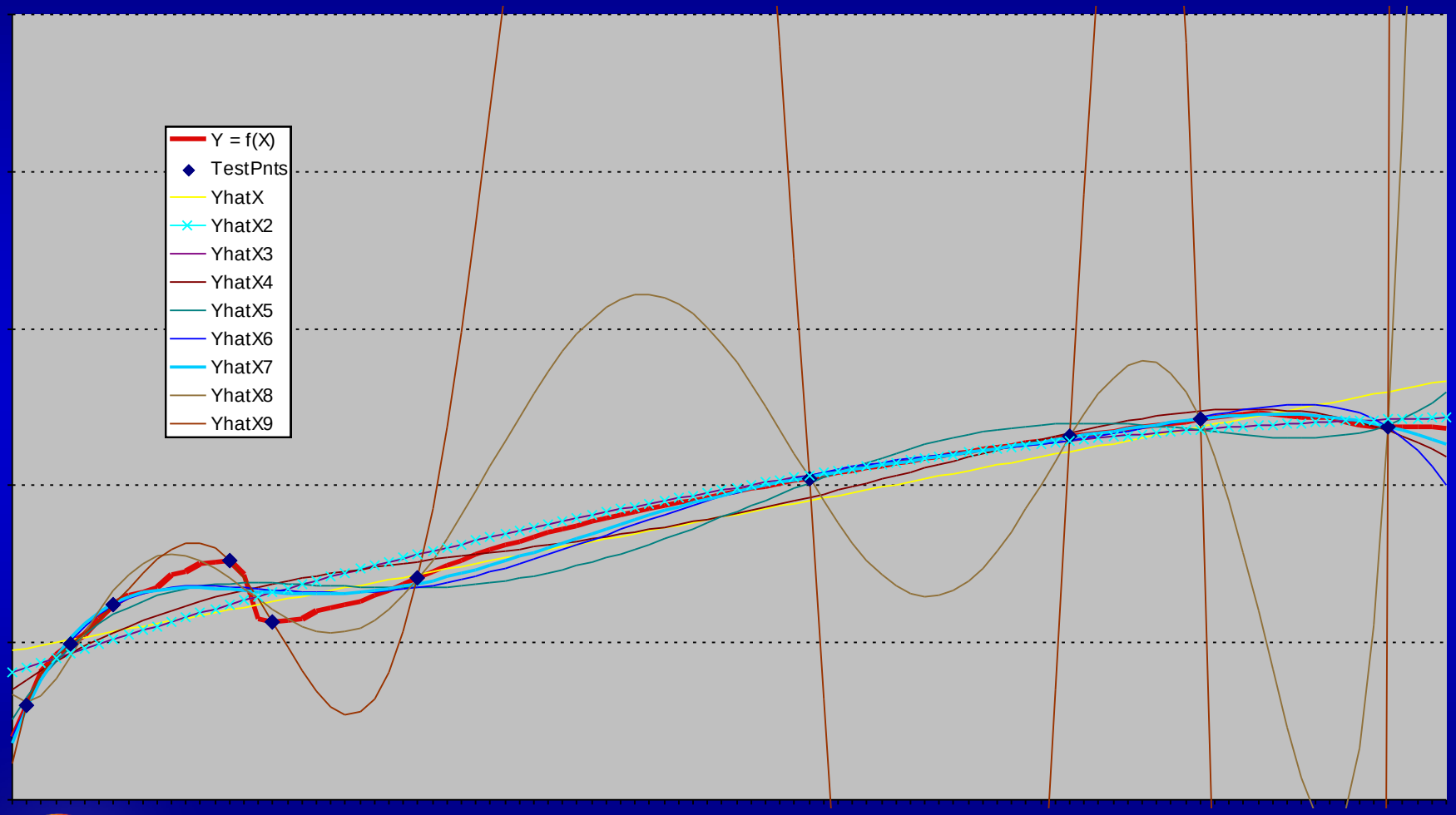
- Only *out-of-sample* results matter.
  (Otherwise, a lookup table wins.)
- Cancer detection example: MD Anderson researchers (1993), using neural networks, were surprised to find that longer training (week vs. day) led to only slightly improved training results, and much worse evaluation results. (They had *overfit* their model.)
- ***Resampling*** is the best defense. (Also known as bootstrap, cross-validation, jackknife, leave-one-out...) It is an *essential* tool. Traditional significance tests are too weak when the model structure is part of the search. Resampling simulations answer: "How likely was that result arrived at by chance?"

Regression error vs. #parameters



Regression Parameters

Overfit models generalize poorly



New Idea: Target Shuffling
can measure the "vast search effect"

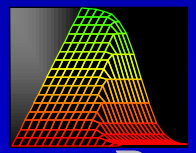
1) Break the link between target, Y, and features, $\underline{X}$ by shuffling Y to form Y_s.
2) Model new $Y_s \sim f(\underline{X})$
3) Measure the quality of resulting (random) model
4) Repeat to build distribution (of random models)
5) True model performance can be measured against this distribution. (The best (or mean) shuffled model can be the baseline for comparison.)

(This capability added to Salford System's CART at our suggestion)

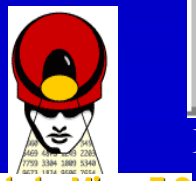
3. Rely on One Technique

- "To a little boy with a hammer, all the world's a nail."
  For best work, you need a whole toolkit.
- Always compare your results to those of a conventional method (e.g., linear regression, or linear discriminant analysis).
- Study: In refereed Neural Network journals, over a 3 year period, 5/6 of the articles made mistake 2 or 3; only 1/6 both tested their model on unseen data and compared it to a baseline technique.
- Not checking other methods leads to blaming/crediting the *algorithm* for the results. But, it's unusual for the modeling technique to make a big difference, compared to feature creation, complexity control, etc.
  - Best: Use a handful of good tools. Each adds only 5-10% effort.

Data Mining Products

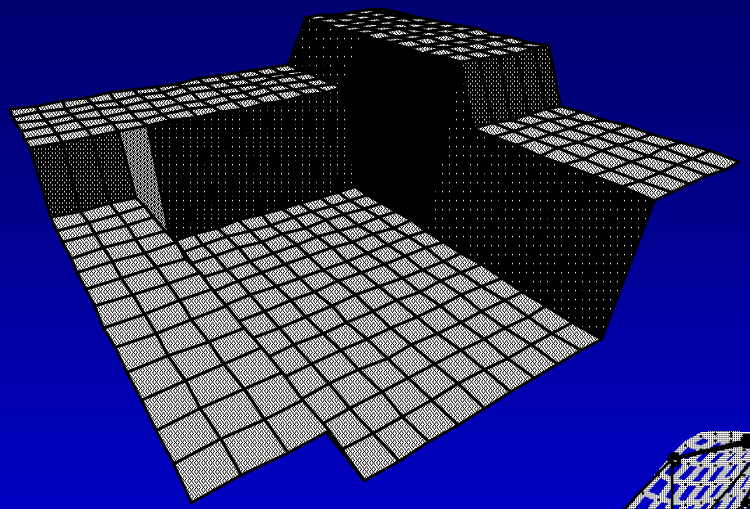


PREDICTIVE DYNAMIX

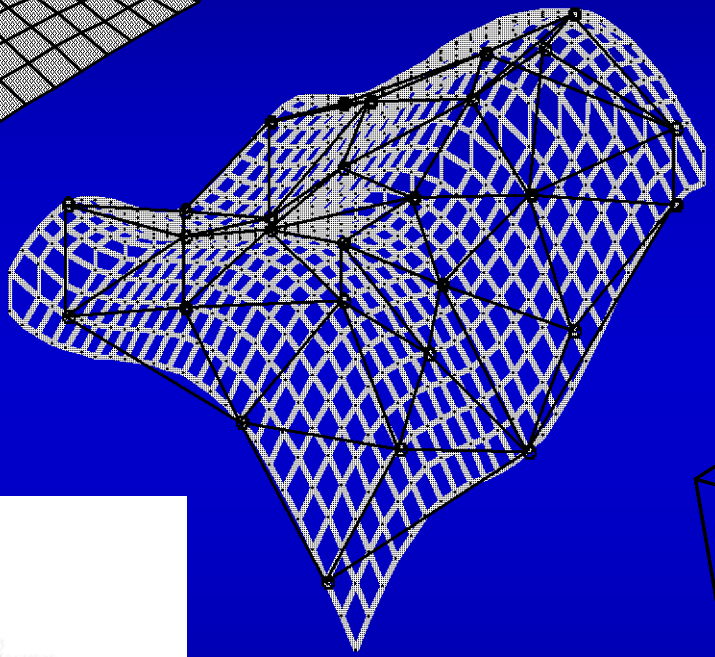


Model 1

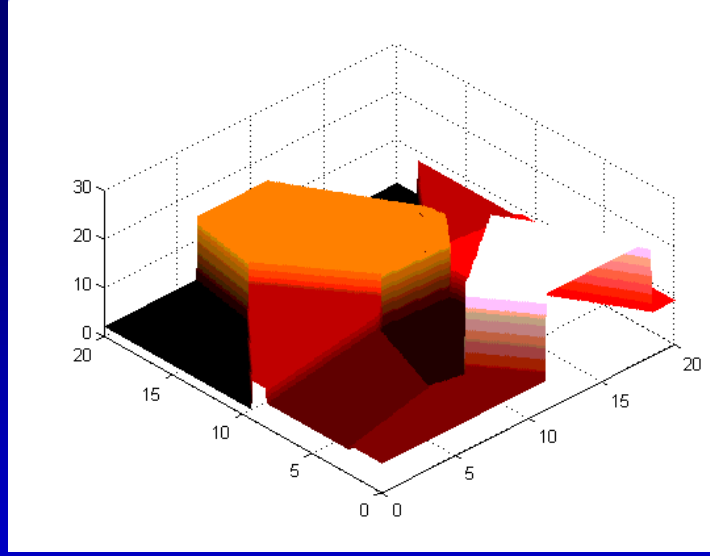
Elder Research, Inc.
Data Mining & Pattern Discovery



Decision Tree

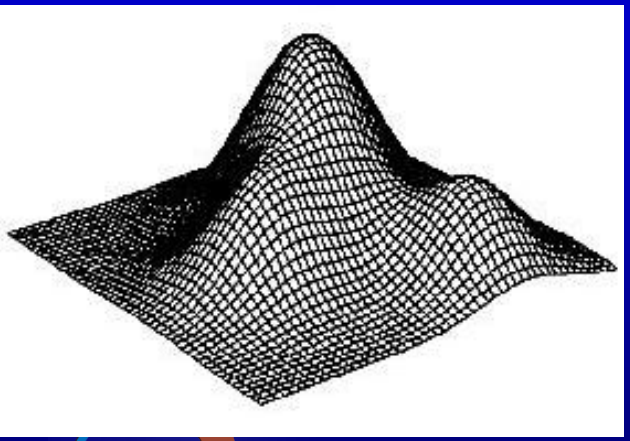


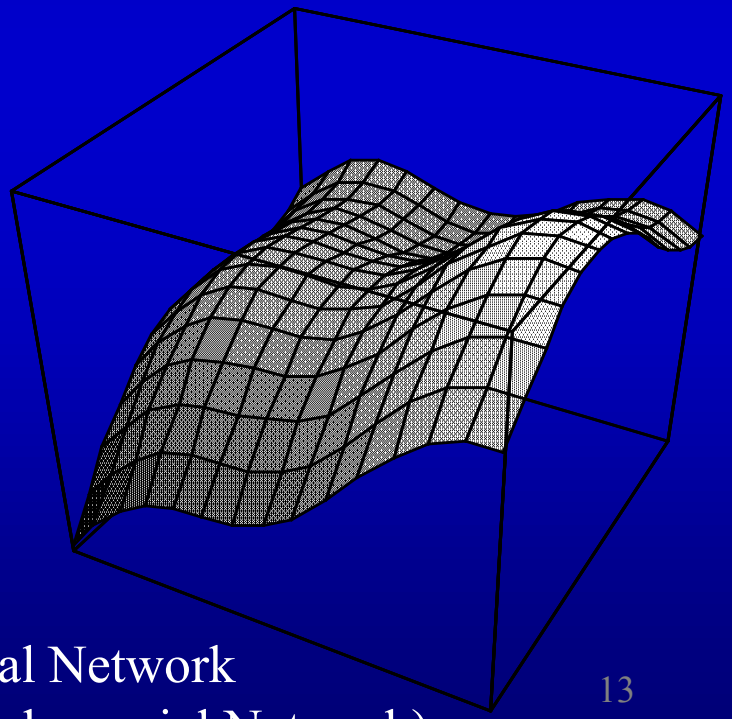
Delaunay Triangles



Nearest Neighbor

Kernel

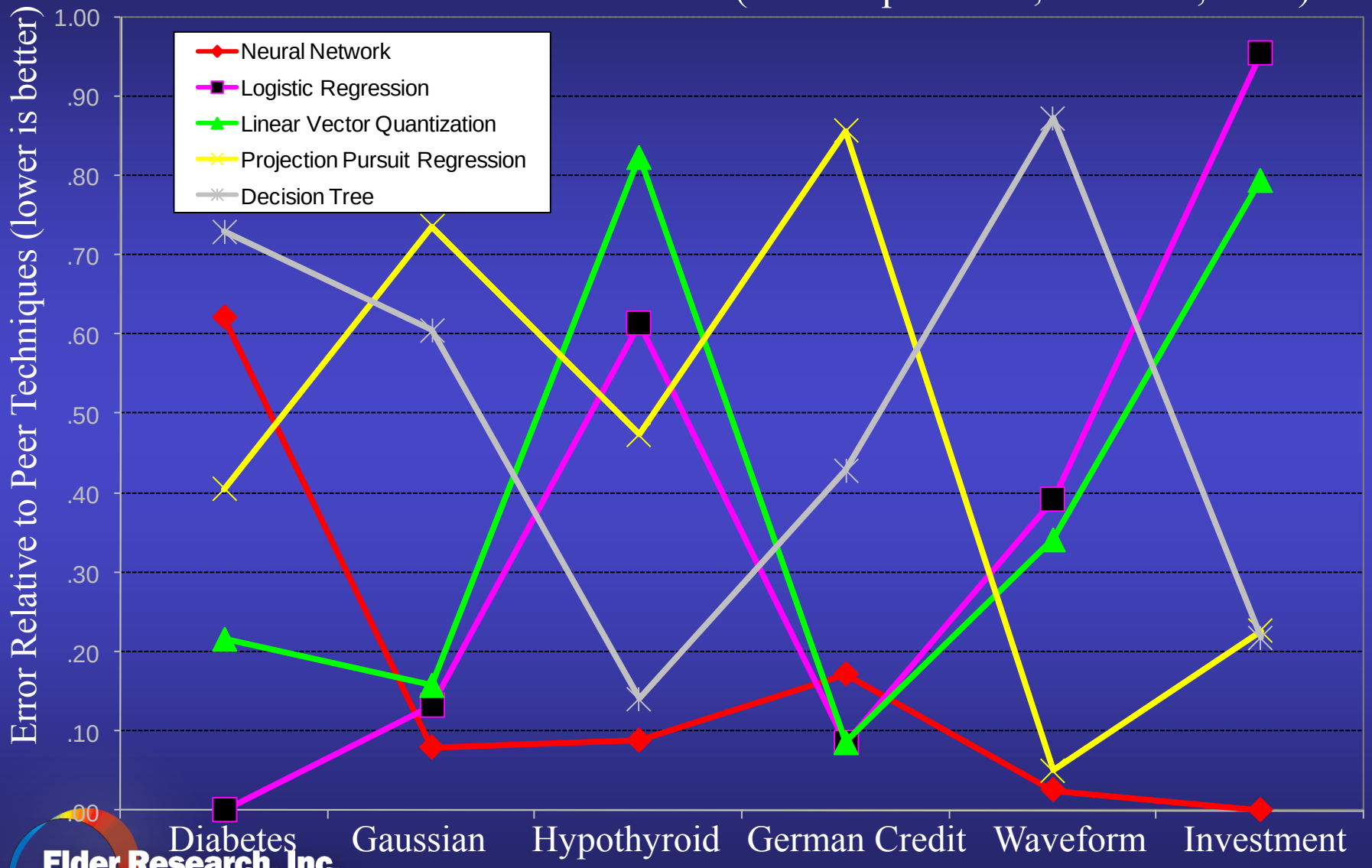




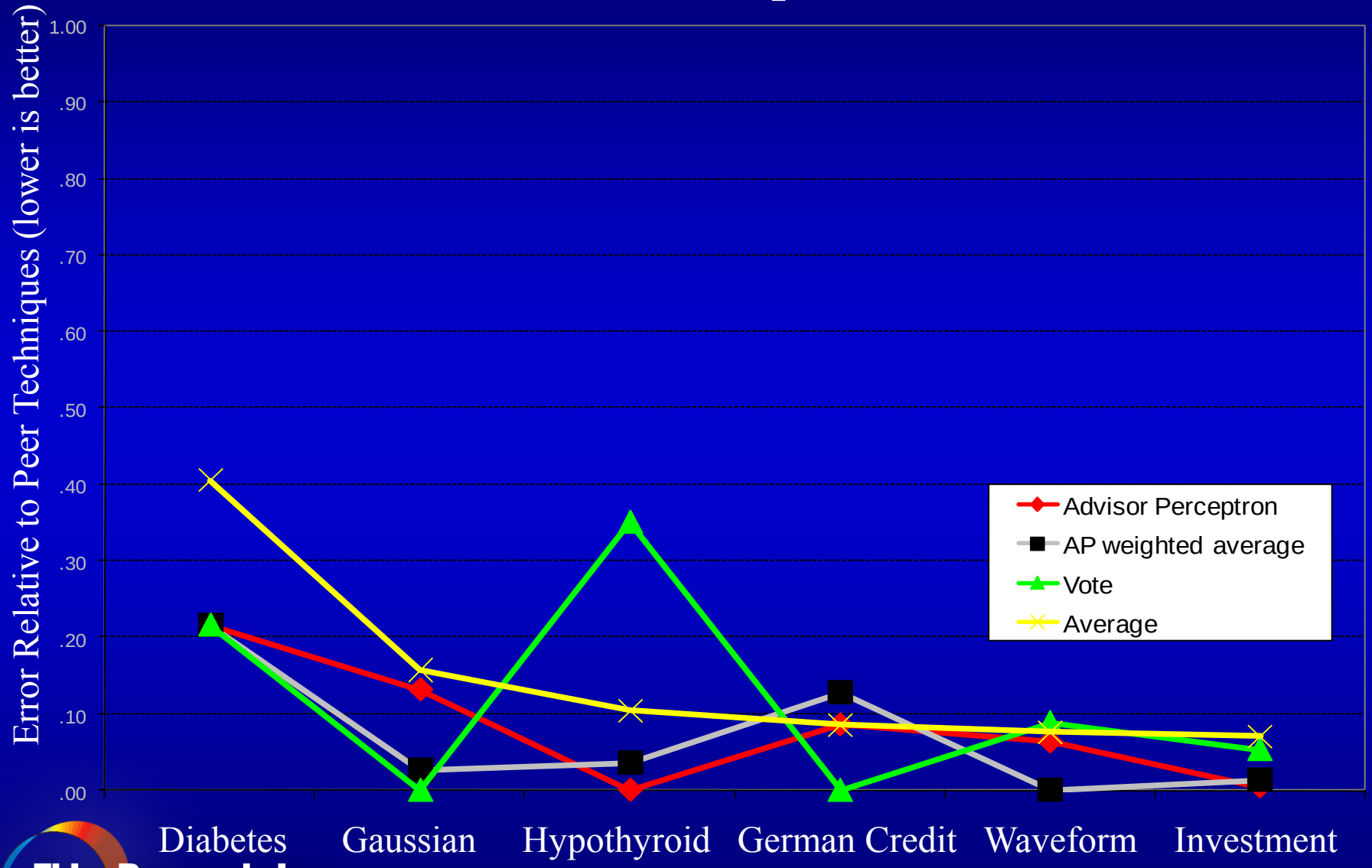
Neural Network
(or Polynomial Network)

Relative Performance Examples: 5 Algorithms on 6 Datasets

(with Stephen Lee, U. Idaho, 1997)



All Ensemble Methods Improve Performance



4. Ask the Wrong Question

4a. Project Goal: Aim at the right target

- Fraud Detection (Positive example!) (Shannon Labs work on Int'l calls): Didn't attempt to classify fraud/nonfraud for general call, but characterized normal behavior for each account, then flagged outliers.
  -> A brilliant success.

4b. Model Goal: Get the computer to "feel" like you do
[e.g., employee stock options]

- Most researchers are lulled into the realm of squared error by its convenience (mathematical beauty). But ask the computer to do what's most helpful for the system, not what's easiest for it.
  [Stock market ex.]

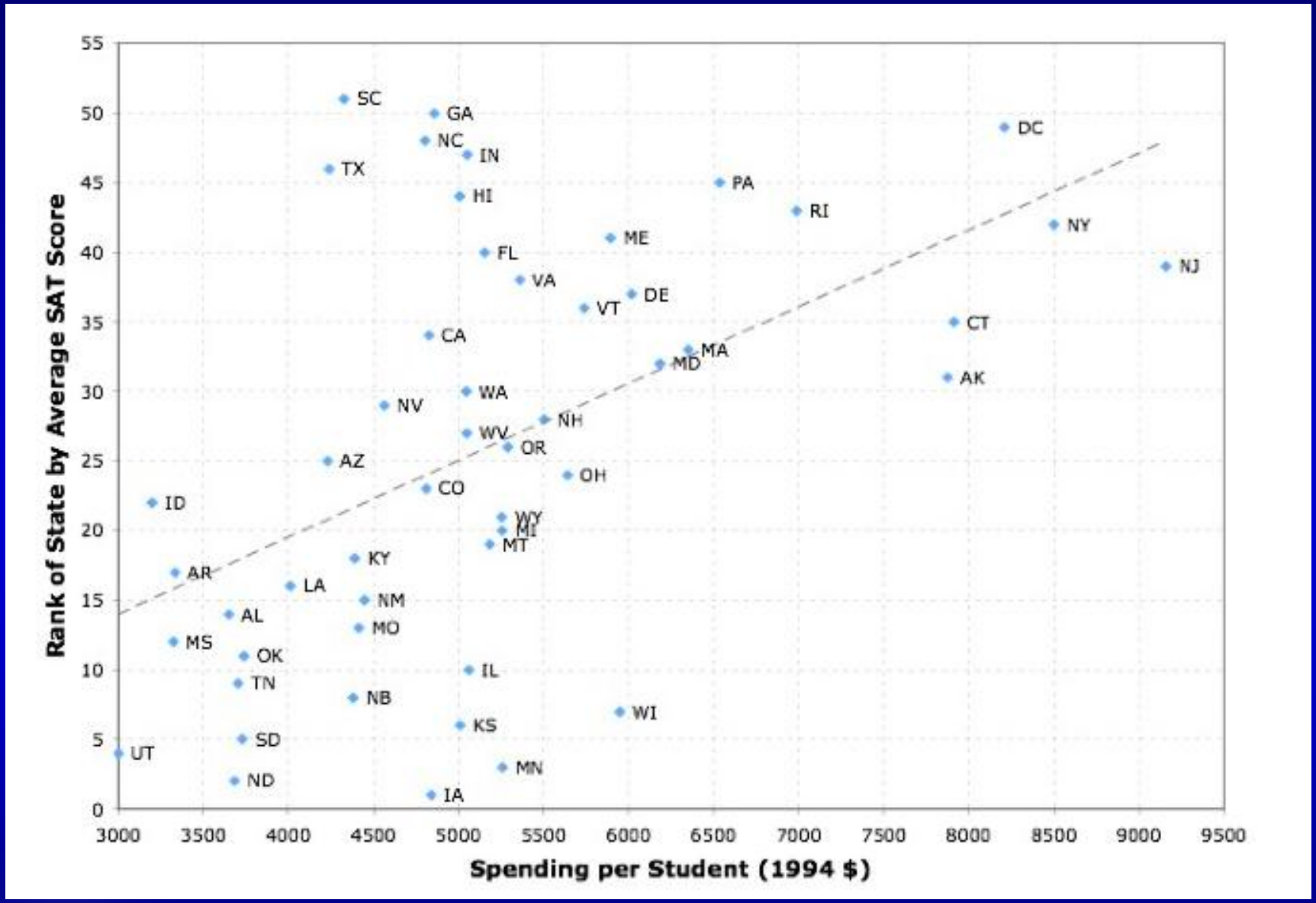
5. Listen (only) to the Data

5a. Opportunistic data:

- [School funding ex.] *Self-selection*. Nothing inside the data protects analyst from significant, but wrong result.

5b. Designed experiment:

- [Tanks vs. Background with Neural networks]: Great results on out-of-sample portion of database. But found to depend on random pixels (Tanks photographed on *sunny* day, Background only on *cloudy*).



Elder Research, Inc.
Data Mining & Pattern Discovery

6. Accept Leaks from the Future

- Forecasting example: Interest rate at Chicago Bank. Neural net 95% accurate, but output was a candidate input.
- Hedge Fund example: Strategy turned out to be moving average of 3 days, but centered on *today*.
- Look for (and remove) variables which work too well. Insurance Example: code associated with 25% of purchasers turned out to describe the type of *cancellation*.
- Date-stamp records when storing in Data Warehouse, or Don't overwrite old value unless archived.
- Survivor Bias [financial ex.]

7. Discount Pesky Cases

- Outliers may be killing results (ex: decimal point price error), or be a discovery (ex: Ozone hole), so examine carefully.
- The best phrase in research isn't "Aha!", but "That's odd…"
- Internal inconsistencies in the data may reveal problems with the flow of information and reveal a larger business problem.
- Direct Mail example: Persisting in hunting down oddities found errors by Merge/Purge house, and was a major contributor to doubling sales per catalog.
- Visualization can cover a multitude of assumptions.

4 Series: (X,Y_1) (X,Y_2) (X,Y_3) (X_4,Y_4)

X	Y_1	Y_2	Y_3	X_4	Y_4
10	8.04	9.14	7.46	8	6.58
8	6.95	8.14	6.77	8	5.76
13	7.58	8.74	12.74	8	7.71
9	8.81	8.77	7.11	8	8.84
11	8.33	9.26	7.81	8	8.47
14	9.96	8.10	8.84	8	7.04
6	7.24	6.13	6.08	8	5.25
4	4.26	3.10	5.39	19	12.50
12	10.84	9.13	8.15	8	5.56
7	4.82	7.26	6.42	8	7.91
5	5.68	4.74	5.73	8	6.89

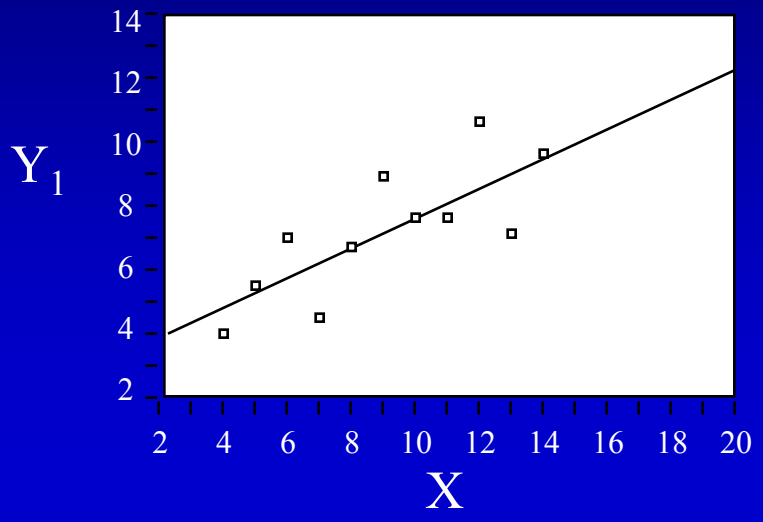
$\rho_{xy} = 0.85$

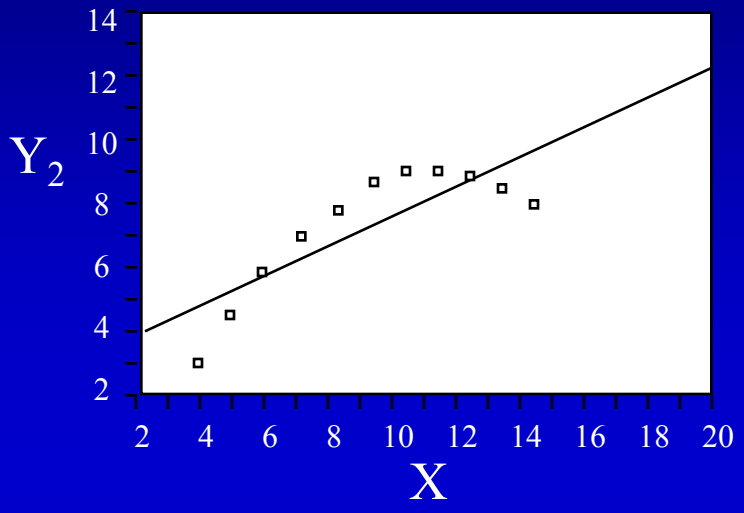
$\hat{y}_{LS} = 3 + 0.5x$

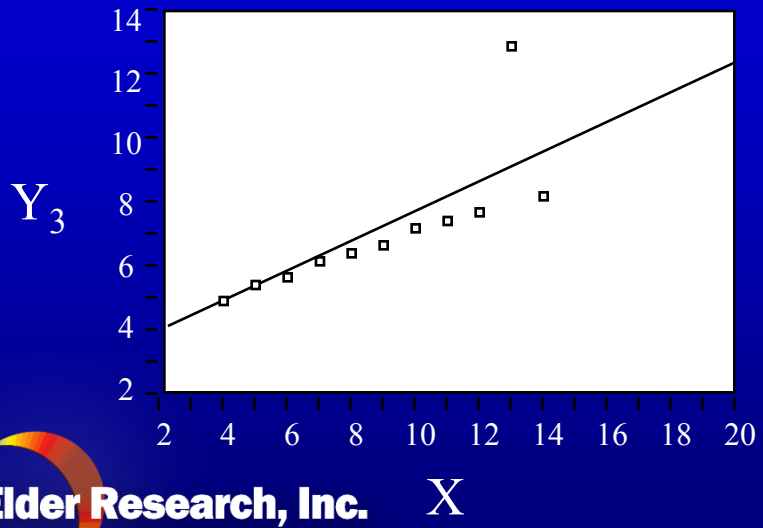
MSE = 1.25

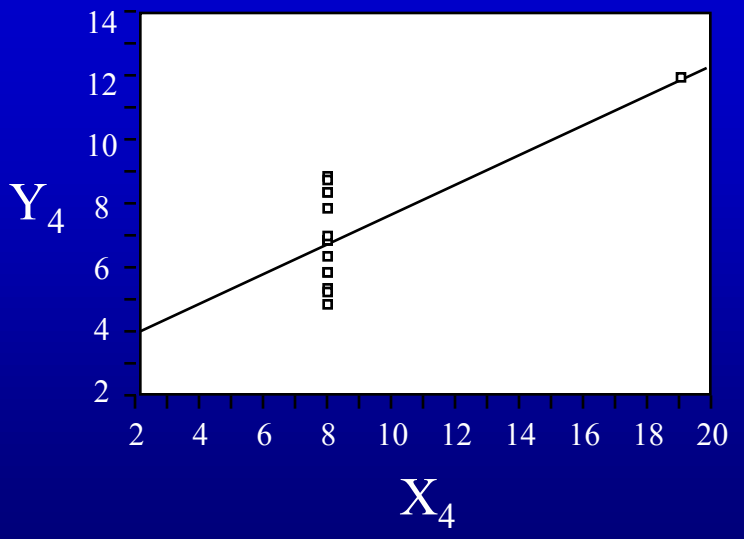
$R^2 = 0.67$

Anscomb's Quartet (1973, *American Statistician*)









Elder Research, Inc.
Data Mining & Pattern Discovery

8. Extrapolate

- Tend to learn too much from first few experiences.
- Hard to "erase" factoids after an upstream error is discovered.
- *Curse of Dimensionality*: low-*d* intuition is useless in high-*d*.
- *Philosophical*: Evolutionary Paradigm:

  Believe we can start with pond scum (pre-biotic soup of raw materials)
  + *zap* + time + chance + differential reinforcement -> a critter.

  (e.g., daily stock prices + MARS -> purchase actions,
  or pixel values + neural network -> image classification)

  Better paradigm is *selective breeding*:

  mutts + time + directed reinforcement -> greyhound

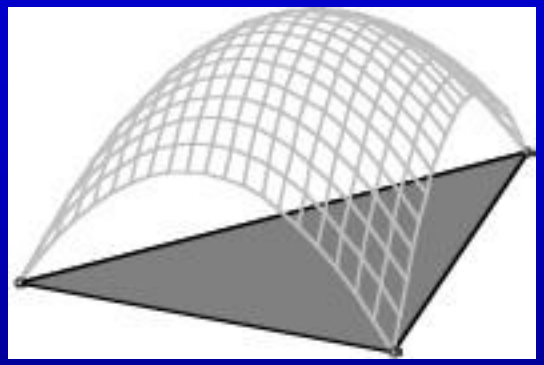
  Higher-order features of data + domain expertise are essential

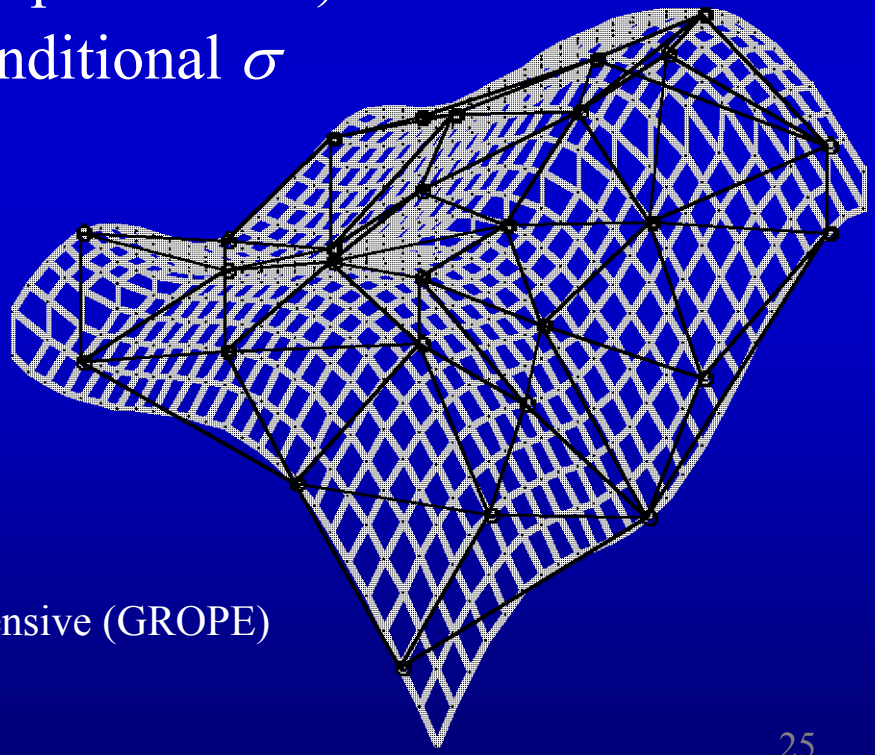
"Of course machines can think. After all, humans are just machines made of meat."
- MIT CS professor

Human and computer strengths are more complementary than alike.

9. Answer Every Inquiry

- "Don't Know" is a useful model output state.
- Could estimate the *uncertainty* for each output (a function of the number and spread of samples near X).
  Few algorithms provide a conditional σ
  with their conditional μ.





Global R^d Optimization when Probes are Expensive (GROPE)

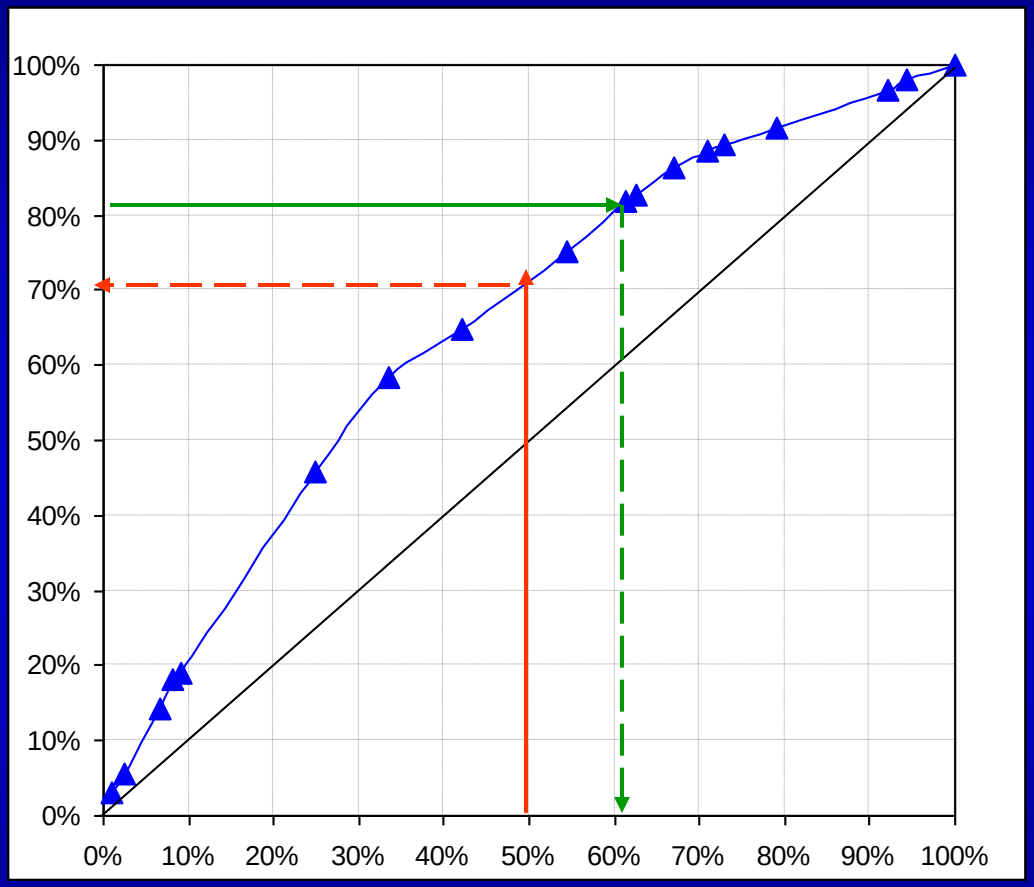
10. Believe the Best Model

- Interpretability not always necessary.
  Model can be useful without being "correct" or explanatory.
- Often, particular variables used by "best" model (which barely won out over hundreds of others of the millions (to billions) tried, using a score function only approximating one's goals, and on finite data) have too much attention paid to them. (*Un-interpretability* could be a virtue!).
- Usually, many very similar variables are available, and the particular structure of the best model can vary chaotically. [Polynomial Network Ex.] But, structural similarity is different from functional similarity. (Competing models often look different, but act the same.)
- Best estimator is likely to be an *ensemble* of models.

Using Lift Charts

2a. Or, Set desired response

1b. Note expected response



1a. Set investigation limit

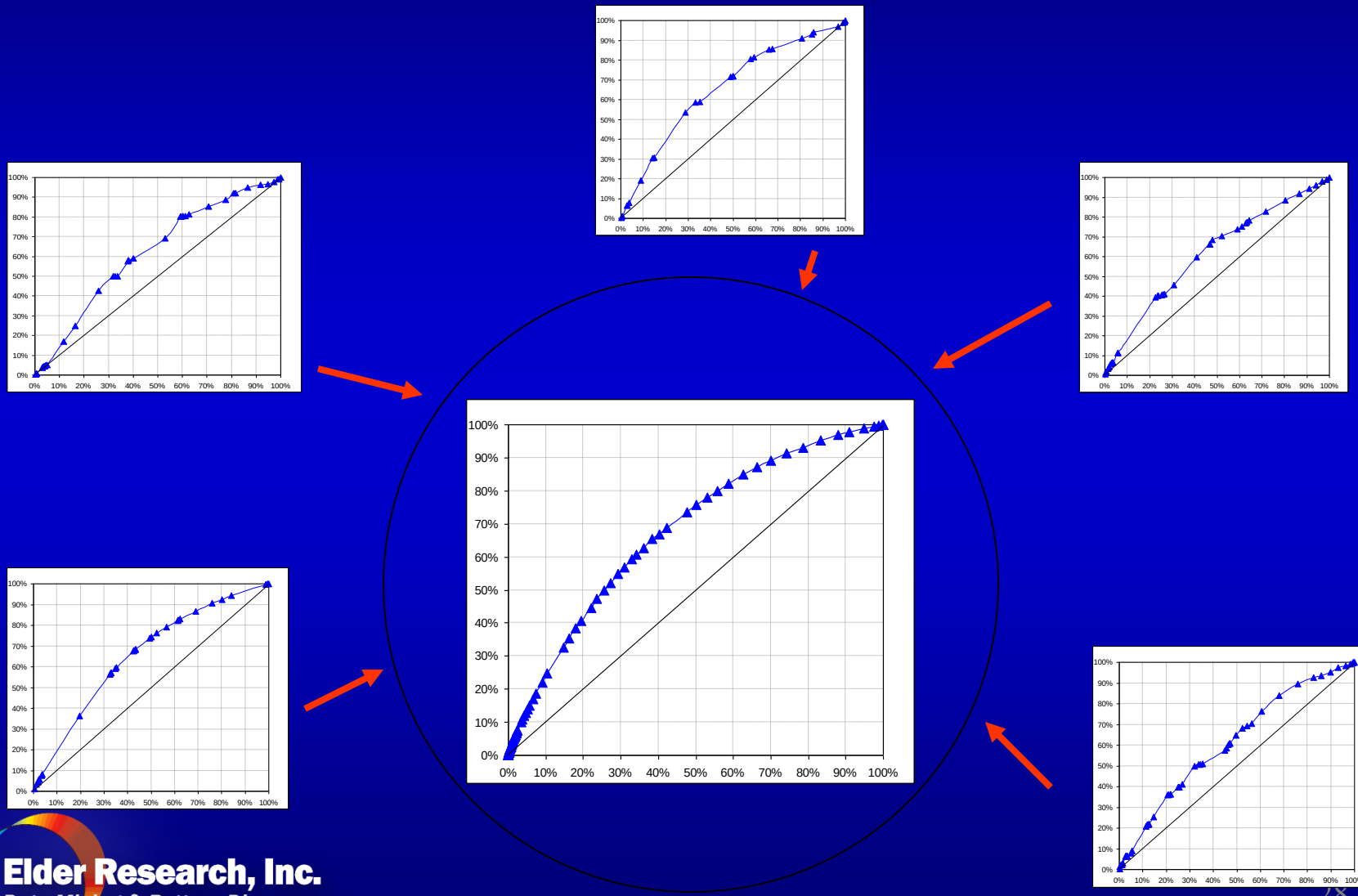
2b. And note work requirements

Prospects Ordered by Response Probability

Elder Research, Inc.
Data Mining & Pattern Discovery

Bundling (Ensembling) 5 Trees

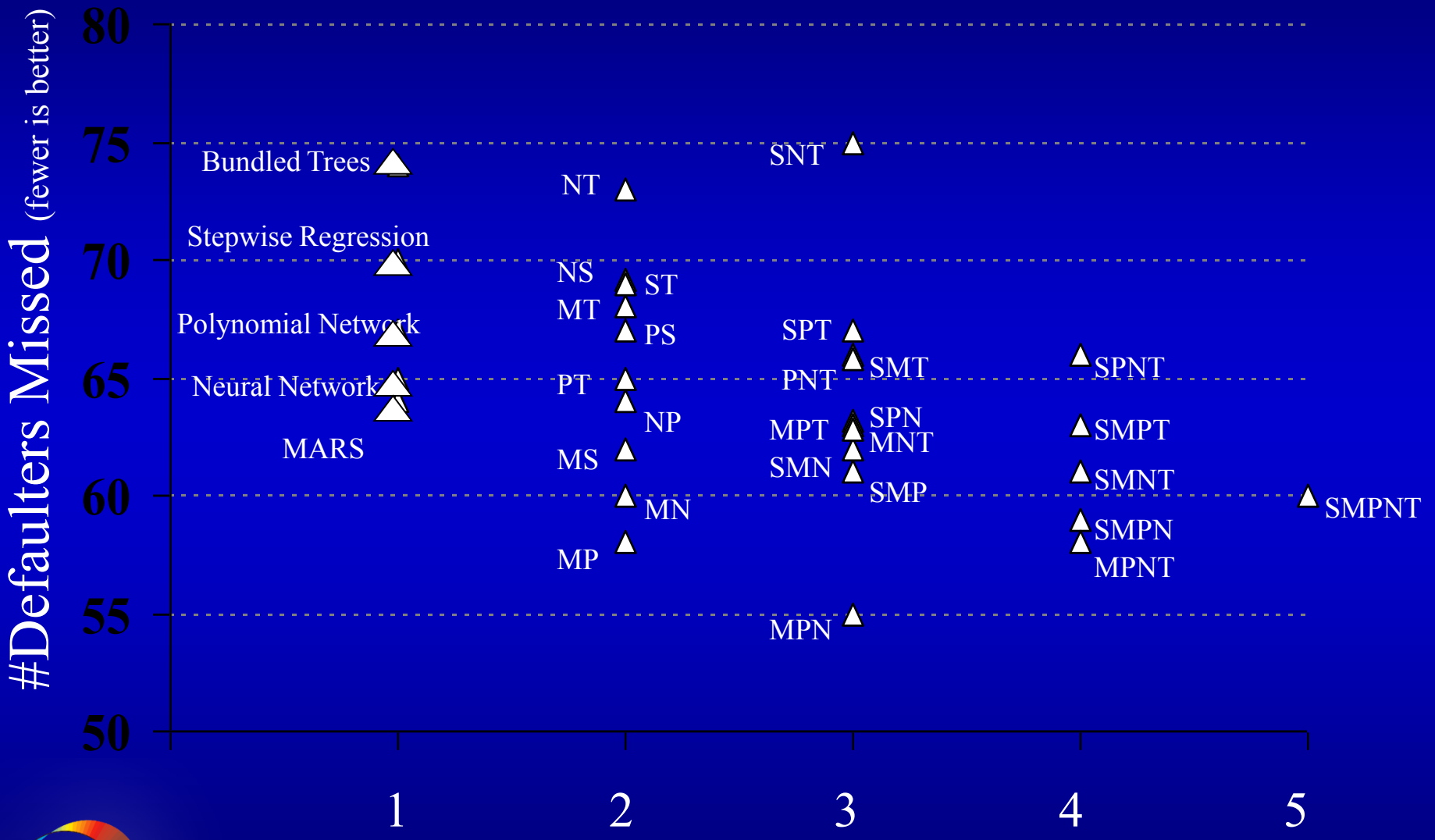
Improves lift, smoothness, and number of decision points



Application Example: Credit Scoring
(Elder Research 1996-1998)

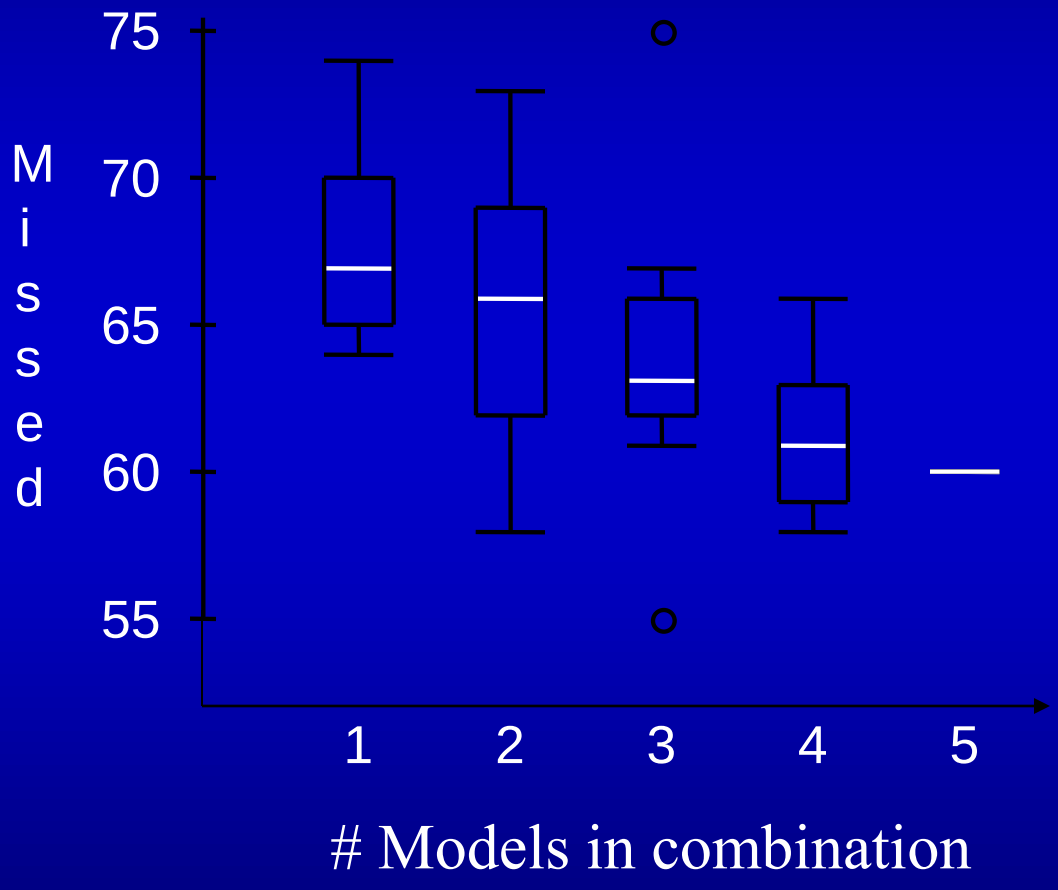
- After 2 years experience, label credit accounts:
  0 (good), 1 (*default* = 90 days late at least once).
- Create models to forecast this outcome
  using only information known at time of credit application.
- Use several (here, 5) different algorithms,
  all employing the same candidate model inputs.
- Rank-order accounts:
  - Give highest-risk value a rank of 1, second highest 2, etc.
  - For bundling, combine model ranks (not estimates) into a new consensus estimate (which is again ranked)
- Report number of defaulting accounts missed (in top portion)

Credit Scoring Model Performance



Elder Research, Inc.
Data Mining & Pattern Discovery

Median (and Mean) Error Reduced with each Stage of Combination



Fancier tools and harder problems –> more ways to mess up.

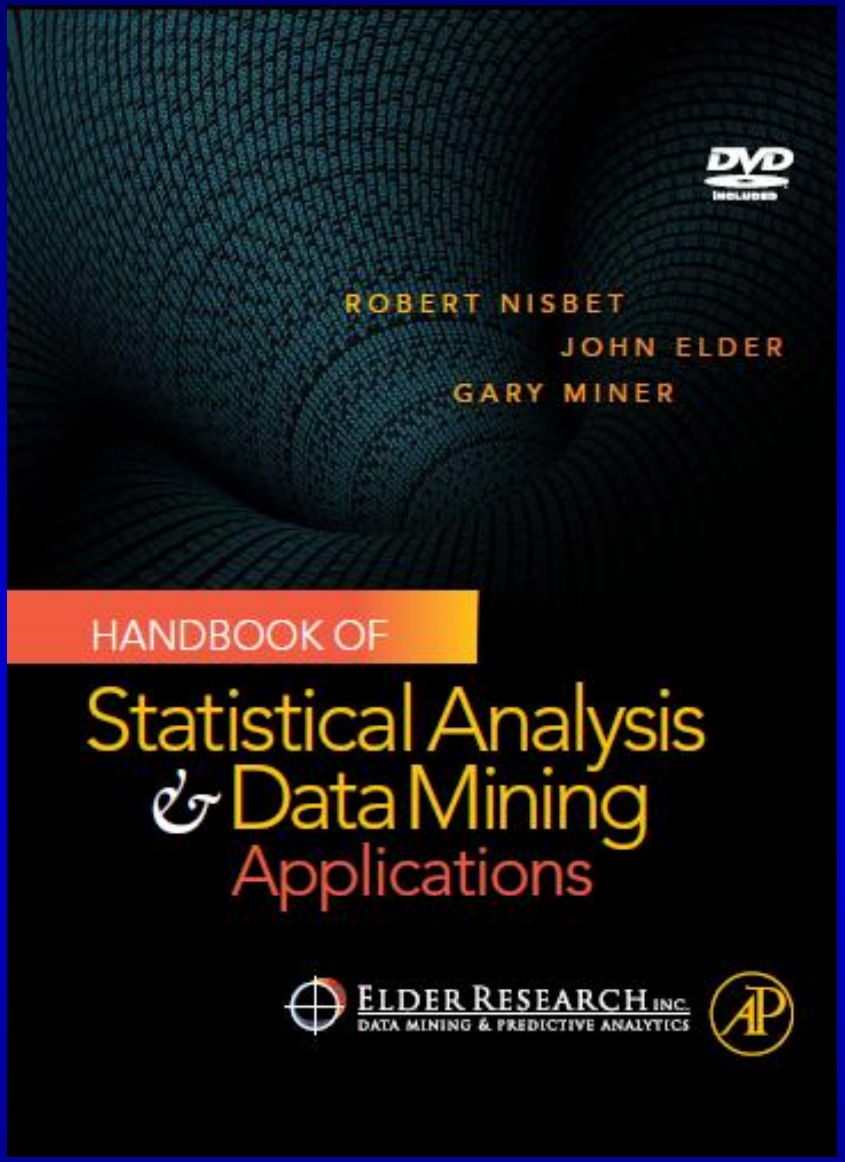
How then can we succeed?

Success <- Learning <- Experience <- Mistakes

(so go out and make some good ones!)

PATH to success:

- **Persistence** - Attack repeatedly, from different angles.
  Automate essential steps. Externally check work.
- **Attitude** - Optimistic, can-do.
- **Teamwork** - Business and statistical experts must cooperate.
  Does everyone *want* the project to succeed?
- **Humility** - Learning from others requires vulnerability.
  Don't expect too much of technology.



www.tinyurl.com/bookERI

John F. Elder IV
Chief Scientist, Elder Research, Inc.

Dr. John Elder heads a data mining consulting team with offices in Charlottesville, Virginia and Washington DC (*www.datamininglab.com*). Founded in 1995, Elder Research, Inc. focuses on Federal, commercial, investment, and security applications of advanced analytics, including text mining, stock selection, image recognition, biometrics, process optimization, cross-selling, drug efficacy, credit scoring, risk management, and fraud detection.

John obtained a BS and MEE in Electrical Engineering from Rice University, and a PhD in Systems Engineering from the University of Virginia, where he's an adjunct professor teaching Optimization or Data Mining. Prior to 14 years at ERI, he spent 5 years in aerospace defense consulting, 4 heading research at an investment management firm, and 2 in Rice's *Computational & Applied Mathematics* department.

Dr. Elder has authored innovative data mining tools, is a frequent keynote speaker, and was co-chair of the 2009 *Knowledge Discovery and Data Mining* conference, in Paris. John's courses on analysis techniques – taught at dozens of universities, companies, and government labs – are noted for their clarity and effectiveness. Dr. Elder was honored to serve for 5 years on a panel appointed by the President to guide technology for National Security. His book on Practical Data Mining, with Bob Nisbet and Gary Miner, was published in May 2009.

John is a follower of Christ and the proud father of 5.

Elder Research, Inc.
Data Mining & Pattern Discovery

M2009 Data Mining Conference

October 26-27 Caesars Palace, Las Vegas

www.sas.com/m2009

Save 30% on Conference Fees!

Webinar attendees are eligible for a 30% discount on conference fees. Reference the discount **DM30** when you register.

If registering online, put **DM30** in the comments field at the bottom of the registration form.

Conference Highlights

- 6 Keynote Speakers
  - Bart Baesens, Katholieke Universiteit Leuven and University of Southampton
  - Michael Berthold, University of Konstanz
  - John Elder, Elder Research, Inc
  - Manfred Krafft, University of Munster
  - Kim Larsen, Charles Schwab & Co.
  - Will Neafsey, Ford Motor Company
- 30+ session talks on a variety of topics
- Visit **www.sas.com/m2009** for a complete list of speakers, abstracts and pre- and post-conference training options.

Questions?

Ask a question by typing your question in the box and clicking "Submit Question"

For more information

John F. Elder IV, PhD

elder@datamininglab.com

Elder Research, Inc.

300 West Main Street, Suite 301

Charlottesville, Virginia 22903

434-973-7673

www.datamininglab.com

concierge@bettermanagement.com

BETTER MANAGEMENT™

Top 10 Data Mining Mistakes

Visionary perspectives for management insights